An analytically solvable model for soccer: further implications of the classical Poisson model

John Fry and Tom Hastings and Jean-Philippe Serbera

School of Computing Mathematics and Digital Technology, Manchester Metropolitan University, Sheffield Management School, Sheffield University, Sheffield Business School, Sheffield Hallam University

June 2017

An analytically solvable model for soccer: further implications of the classical Poisson model

July 2017

Abstract

In this paper we discuss an exactly soluble statistical model for soccer. By taking into account key features of soccer matches (goals are rare, goal-scoring patterns are not well understood) we arrive at a version of the classical Poisson model but with constraints on the expected total number of goals in a game. Closed form expressions are derived for expected scores and match outcomes. We are also able to reconstruct an empirically observed inverse strike-back effect pertaining to teams scoring consecutive goals. We produce analytical results for perfectly competitive soccer leagues where draws are tied to the average number of goals in a game. An empirical application to the 2016 UEFA European Championships is also discussed.

Keywords: Forecasting; Gaming; Sports; Stochastic Processes

1 Introduction

Soccer is the world's most popular sport (Constantinou and Fenton, 2015). Mass participation rates, soccer is played by an estimated 250m people across over 200 countries, are accompanied by a global TV audience of billions. Viewing figures for soccer's World Cup exceed even those of the Olympics. The public's fasicnation with soccer is reinforced by intense interest in sports modelling in general (see e.g. Percy, 2009). Allied to the above the quantitative modelling of soccer is particularly in vogue (see e.g. Anderson and Sally, 2015; Baker and McHale, 2015;

Kuyper and Szymanski, 2014). As such our contribution is important for a number of theoretical and practical considerations beyond considerable public interest.

The theoretical importance of our contribution is as follows. Our model reflects both a rich history of Poisson modelling in football (Maher, 1982; Keller, 1994) and the fact that soccer's low scoring nature and high numbers of draws stands in sharp contrast to North American sports (Maher, 1982; Wright, 2009). Soccer's rigid league structures and well-defined promotion, relegation, tournament qualification and winning targets also mean it is especially amenable to the kind of analysis considered here. Further, our model abstracts from key features of soccer matches such as the fact that goal-scoring patterns are not very well understood (Kuper and Szymanski, 2014) amid other stylised empirical facts (Dixon and Robinson, 1998).

The practical importance of our contribution is three-fold. Firstly, we derive analytical expressions for individual scores and match outcomes. This includes modifications to account for complications like extra-time and penalties which may have some relevance to sports-betting applications (see e.g. Fitt et al., 2006). We consider both a Mean-Variance approximation and a higher-order Mean-Variance-Kurtosis approximation that leads to improved results in empirical applications. Secondly, in a simplified version of the full model we undertake theoretical analysis of perfectly competitive soccer leagues. Though an obvious simplification this enables us to derive realistic performance benchmarks for several major competitions. We can recreate versions of several rules of thumb that are well known throughout football. Inter alia our model suggests teams target 41 points and 49 points to avoid relegation from the English Premier League and the English Championship respectively (see the third section of this paper). Thirdly, we apply the full version of our model to the recent UEFA 2016 European Championship and show that some of the supposedly shock results are not the complete surprise they may at first appear. We can thus demonstrate that our model is easy to apply in practice and can lead to conclusions that are far from trivial. A further implication is that many fans and pundits fundamentally under-estimate soccer's inherent unpredictability. Our model makes efficient use of the vast amount of historical information available in terms of the average number of goals scored per match. This simple statistic both reflects the co-evolution of increasingly advanced offensive and defensive strategies (Anderson and Sally, 2013) and, in showing marked variation

between elite and non-elite leagues, is itself highly informative about the nature of individual competitions. For major competitions the typical number of goals per game is around 2.4-2.75 per game with average values of less than 2 goals per game and more than 3.5 goals per game thought to be exceptional and to usually be indicative of lower overall standards of play.

The layout of this paper is as follows. The second section outlines the model used and highlights the key analytical results obtained. The third section discusses perfectly competitive soccer leagues – an idealised setting that nonetheless has some interesting implications for the setting of performance benchmarks in elite soccer leagues. The fourth section discusses an empirical application to the 2016 UEFA European Championship finals. The final section concludes and discusses further work. A mathematical appendix is included at the end of the paper.

The model

Suppose that, as a simplification, the number of goals scored in a match is distributed as Poisson(λ) and that if a goal is scored then it is scored by team X with probability p and by team Y with probability $1-p$. Thus, in this model goals are rare events and occur at random. This shares qualitative features of football matches discussed in Anderson and Sally (2013) and Kuper and Szymanski (2014) with the nature of real goal-scoring patterns not necessarily very well understood. Our model thus abstracts key details like overall team strengths and soccer's low-scoring nature. Further modifications to take into account the effect of home advantage are discussed in the fourth section of this paper. More detailed effects such as short-term form, time of the season, managerial changes, fatigue, psychological judgements, bookmakers information and subjective judgements have been discussed in recent applied work (see e.g. Owen, 2011; Constantinou et al., 2012; Constantinou and Fenton, 2017) but have not been included in our theoretical model here.

Set up in this way our model arises as a special case of the classical model of Maher (1982) but with an additional constraint concerning the expected total number of goals scored in a game. As discussed above this reflects both the wealth of statistical information that has only relatively recently been made available combined with the important league-specific information that this simple statistic encodes.

Proposition 1 *Let $(X = x, Y = y)$ denote the event that Team X scores x goals and Team Y scores y goals. We have that*

$$Pr(X = x, Y = y) = \frac{e^{-\lambda}(\lambda p)^x(\lambda(1-p))^y}{x!y!}. \tag{1}$$

Proof.

$$Pr(X = x, Y = y) = \frac{e^{-\lambda}\lambda^{x+y}}{(x+y)!}.\frac{(x+y)!p^x(1-p)^y}{x!y!} = \frac{e^{-\lambda}\lambda^{x+y}p^x(1-p)^y}{x!y!}.$$

□

As a simple corollary we have:

Proposition 2 (Expected score) *We have the following results for the expected final score:*
(i) The expected final score is given by

$$E[X] = \lambda p,\ E[Y] = \lambda(1-p).$$

(ii) If the score is $X = x$, $Y = y$ after M minutes the expected final score is given by

$$E[X] = x + \left(1 - \frac{M}{90}\right)\lambda p,\ E[Y] = y + \left(1 - \frac{M}{90}\right)\lambda(1-p).$$

In addition to individual match scores we can write down and solve the probabilities of individual match outcomes:

Proposition 3 (Probability of match outcomes.) *We have the following results for overall match outcomes*
(i) The probability of a drawn is given by $e^{-\lambda}I_0(2\lambda\sqrt{p(1-p)})$,
(ii) The probability that Team X wins is given by $Q_0(\sqrt{2\lambda p}, \sqrt{2\lambda(1-p)})$,
(iii) The probability that Team Y wins is given by $Q_0(\sqrt{2\lambda(1-p)}, \sqrt{2\lambda p})$,
where where I_k denotes the modified Bessel function of the first kind (Abramowitz and Stegun, 1968) and $Q_0(\cdot)$ denotes the Marcum Q-function (Nuttall, 1975).

Proof.

See the Appendix.

The probabilities calculated in Proposition 3 can be evaluated numerically using simulation techniques (see e.g. Weinberg, 2006). However, using the recursion formula in Annamalai and Tellambura (2008) it can be shown that the Marcum Q-function can be written as

$$Q_0(\alpha, \beta) = 1 - F_{2,\alpha^2}(\beta^2) - e^{-\frac{\alpha^2+\beta^2}{2}} I_0(\alpha\beta),$$

where $F_{2,\alpha^2}(\cdot)$ denotes the Cumulative Distribution Function (CDF) of the non-central χ^2 distribution with 2 degrees of freedom and non-centrality parameter α^2. Motivated by empirical implied probabilities obtained from bookmakers' odds we have the following extension to Proposition 3 for one-off matches in knock-out competitions:

Proposition 4 (Outcomes in one-off knock-out matches.) *Assume that in a penalty shoot-out each team is equally likely to win. Suppose a knock-out game goes to extra-time*
(i) The conditional probability that team X wins after extra time (aet) is given by

$$Pr\left(X \text{ aet}\right) = \frac{1}{2} Q_0\left(\sqrt{\frac{2\lambda p}{3}}, \sqrt{\frac{2\lambda(1-p)}{3}}\right) + \frac{1}{2}\left(1 - Q_0\left(\sqrt{\frac{2\lambda(1-p)}{3}}, \sqrt{\frac{2\lambda p}{3}}\right)\right). \tag{2}$$

(ii) The conditional probability that team Y wins after extra time (aet) is given by

$$Pr(\, Y \text{ aet}) = \frac{1}{2} Q_0\left(\sqrt{\frac{2\lambda(1-p)}{3}}, \sqrt{\frac{2\lambda p}{3}}\right) + \frac{1}{2}\left(1 - Q_0\left(\sqrt{\frac{2\lambda p}{3}}, \sqrt{\frac{2\lambda(1-p)}{3}}\right)\right). \tag{3}$$

(iii) The probability that team X wins the knockout match is given by

$$Pr\left(X \text{ wins}\right) = Q_0\left(\sqrt{2\lambda p}, \sqrt{2\lambda(1-p)}\right) + e^{-\lambda} I_0(2\lambda\sqrt{p(1-p)}) Pr(X \text{ aet}). \tag{4}$$

(iv) The probability that team Y wins the knockout match is given by

$$Pr\left(X \text{ wins}\right) = Q_0\left(\sqrt{2\lambda(1-p)}, \sqrt{2\lambda p}\right) + e^{-\lambda} I_0(2\lambda\sqrt{p(1-p)}) Pr(\, Y \text{ aet}). \tag{5}$$

Proof

See the Appendix.

Dixon and Robinson (1998) find that the following stylised empirical facts are associated with soccer matches.

1. The scoring rate increases as soon as the first goal is scored.
2. The scoring rate depends on the score.
3. The scoring rate for both teams is non-decreasing over time.
4. The inverse strike-back effect. It is not the case that a team is never more vulnerable than after they have just scored a goal. The opposite effect is more likely.
5. If the home team is leading the home and away goalscoring rates generally decrease and increase respectively.

Whilst our simple model is unable to account for time-varying scoring rates our model is able to provide a theoretical description of the empirically observed inverse strike-back effect.

Proposition 5 (The inverse strike-back effect.) *The team that has just scored is also more likely to score the next goal.*

Proof

See the Appendix.

Perfectly competitive sports leagues

The study of competitive balance in sports leagues has a long history dating back to the work of Rottenberg (1956). Analysing the competitive balance of national leagues is of enduring interest (Jessop, 2006; Lee and Fort, 2012). Moreover, evidence suggests that elite soccer teams are already operating at very high levels of efficiency (Espitia-Escuer and García-Cebrián, 2010) and suggests that such leagues may be becoming increasingly competitive. As an idealised model for a perfectly competitive sports league we take $p = 1 - p = 1/2$ in the above. Inter alia this mirrors the way in which in finance (via the Efficient Markets Hypothesis) unpredictability

emerges as a consequence of very high levels of competition (see e.g. Campbell et al., 1997). It follows from Proposition 3 that in this case we have

$$\begin{aligned} Pr(\text{Win}) &= Pr(\text{Lose}) = \frac{1 - e^{-\lambda} I_0(\lambda)}{2} = w, \\ Pr(\text{Draw}) &= e^{-\lambda} I_0(\lambda) = 1 - 2w. \end{aligned} \tag{6}$$

A similar formulation and related formulae can be found in Keller (1994). Here, we link this approach to a theoretical model for perfect competition in Cain and Haddock (2006) as the probability of a draw can be explicitly linked to the typical number of goals in a game. This suggests that around a quarter of all games will end in a draw (see the next subsection). This is in marked contrast for classical models of competitive balance in North American competitions which, excluding the possibility of draws, can be obtained for $w = 1/2$. Using the model we can both refine the discussion of relegation in Gandy (2016) and provide extensions covering various different performance objectives. Inter alia our approach may also enable us to compare and contrast the performance of teams from different eras (see e.g. Baker and McHale, 2015).

Consider a soccer league of n teams who play each other home and away once only. Each team thus plays $2(n-1)$ games over the course of a season. Three points are awarded for a win, one point for a draw and 0 points for a defeat. The objective for team X is to secure enough points over an entire league season so that X defeats a randomly chosen opponent, team Y say, with probability q. For example, the English Premier league consists of $n = 20$ teams. The target to avoid relegation is to defeat at least 3 teams thus finishing 16+1=17th place or higher. This gives $q = (3 \text{ defeated teams})/(19 \text{ rival teams})$. Let X_i denote the number of points scored by team X in game i. We have that $Pr(X_i = 3) = Pr(X_i) = 0 = w$, $Pr(X_i = 1) = 1 - 2w$. It follows that $E[X_i] = 3w + 1 - 2w + 0w = 1 + w$. The expected season total for team X is then given by

$$E[X] = \sum_i E[X_i] = 2(n-1)(1+w). \tag{7}$$

The number of points scored by team X compared to team Y is given by

$$X - Y = \sum_{i=1}^{2n-4} X_i - Y_i + Z_{2n-3} + Z_{2n-2}, \tag{8}$$

where the Z_i represent direct head-to-head matches between the two teams and take the values ± 3 with probability w and 0 with probability $1 - 2w$, the distribution of the X_i and Y_i is as described above and all the random variables in equation (8) are mutually independent. The representation in equation (8) suggests that we can set points targets for Team X to achieve a given objective by setting

$$\text{Target} = E[Y] + F^{-1}_{X-Y}(q), \tag{9}$$

where $F^{-1}(\cdot)$ denotes the inverse CDF of the random variable $X - Y$ and q denotes the probability q that Team X defeats a randomly chosen opponent (see above). We have the following approximation formulae for the end-of-season points targets in perfectly competitive leagues.

Proposition 6 *We have the following approximation formulae for points targets in perfectly competitive leagues:*
i) Mean-Variance (MV) approximation:

$$\textit{Target} = (1 + w)2(n - 1) + \sqrt{w^2[8 - 4n] + w[20n - 4]}\Phi^{-1}(q). \tag{10}$$

ii) Cornish-Fisher Mean-Variance-Kurtosis (MVK) approximation:

$$\textit{Target} = (1 + w)2(n - 1) + \sqrt{w^2[8 - 4n] + w[20n - 4]}\left(\left(1 - \frac{\gamma_2}{8}\right)\Phi^{-1}(q) + \frac{\gamma_2[\Phi^{-1}(q)]^3}{8}\right), \tag{11}$$

where $\Phi^{-1}(\cdot)$ denotes the inverse CDF of the normal distribution and γ_2 denotes the excess kurtosis of the random variable $X - Y$ where,

$$\gamma_2 = \frac{188 + 68n + w(1988n - 5320) + w^2(3600n - 6000) + w^3(1800n - 3000)}{w((8 - 4n)w + (20n - 4))^2}.$$

Proof

See the Appendix.

Equation (10) in Proposition 6 gives the familiar mean-variance approximation. This formula gives reasonable answers for low values of q but begins to get markedly less-accurate as q becomes larger – in effect becoming less accurate higher up the league table. Equation (10) can be obtained as a special case of equation (11) with $\gamma_2 = 0$. Further, as we see in the next subsection, the Cornish-Fisher expansion in equation (11) can be seen to lead to empirically meaningful increases in accuracy.

Numerical example

As an illustration we apply equation (11) to the English Premier league with $n = 20$ teams. In the most recent fully completed season, 2015/16, the average Premier League match had on average 2.7 goals per game. Taking $\lambda = 2.7$ equation (6) thus gives $Pr(\text{Draw}) = 0.258$, $w = Pr(\text{Win}) = Pr(\text{Lose}) = 0.371$. Further calculations give

$$\begin{aligned} \text{Mean} &= (1+w)2(n-1) = 52.095 \\ \text{Variance} &= w^2[8-4n] + w[20n-4] = 136.975 \end{aligned}$$

1. *Winning the league.* This involves defeating all 19 rivals. The method recommended by Hyndman and Fan (1996) suggests taking $q = (19 - 1/3)/(19 + 1/3) = 0.966$. In this case equation (10) gives

$$x = 52.095 + \sqrt{136.975}\Phi^{-1}(0.966) = 73.379.$$

   This suggests that in order to win the League title teams should aim to achieve a minimum total of 74 points. In contrast application of equation (11) suggests a winning target of 77 points. In view of the comparison of teams from different eras (see e.g. Baker and McHale, 2015) the suggestion is that once teams exceed 77 points in a Premier League season they can reasonably be viewed as potential Premier League Champions.

2. *Champions League Qualification.* Qualifying for the Champions League via the domestic football competitions entails finishing in the top four. (Since the 2014/15 season winners of

the Europa League (formerly the UEFA Cup) have also qualified for the Champions League Group stages though we do not account for this second qualification route here.) Since finishing in the top four involves defeating 16 randomly chosen rivals taking $q = 16/19$ in the above and using equation (10) gives

$$x = 52.019 + \sqrt{136.294}\Phi^{-1}(16/19) = 63.835.$$

This suggests that in order to achieve Champions League qualification teams need to aim to achieve a minimum of 64 points. In this case equation (11) also gives the same answer.

3. *Avoiding relegation.* Avoiding relegation entails finishing in the top 17 places in the league or, equivalently, defeating 3 randomly chosen rivals. As such taking $q = 3/19$ and using equation (10) gives

$$x = 52.019 + \sqrt{136.294}\Phi^{-1}(3/19) = 40.354.$$

This suggests that in order to avoid relegation teams need to aim to achieve a minimum total of 41 points. This appears in remarkably good agreement with the well-established rule of thumb that states that Premier League teams typically need to aim to achieve 40 points to avoid relegation. In this case equation (11) also gives the same answer.

As a further illustration of our method Table 1 shows the application of our model throughout the English football pyramid and the suggested points totals required for winning the league championship, achieving automatic promotion, qualifying for the promotion playoffs and avoiding relegation in each case. Once again our projected totals appear in remarkably good agreement with several well-known rules of thumb. In particular, our model suggests that teams need to target around 50 points (to be precise actually 49 points) to avoid relegation from the English Championship. Generally the Mean-Variance (MV) approximation under-estimates the points targets for winning the league and for gaining automatic promotion but the differences between the two methods for qualifying for the playoffs or avoiding relegation are very small. However, the increased accuracy of the Mean-Variance-Kurtosis (MVK) Cornish-Fisher approximation clearly has empirical relevance. Generally, the average number of goals per game increases as

the standard of football decreases although the Championship has fewer goals per game than the Premier League. Similar patterns (where the second tier has a fewer number of goals per game than the top tier league) can also be seen in other European countries and may reflect a more defensive overall style of play. Since the structure of the English football league pyramid is so close to uniform – most divisions consisting of 24 teams – projected points totals are broadly the same throughout. However, there are slight differences lower down the pyramid due to different league structures and declining levels of quality – as reflected by generally higher numbers of goals per game.

Empirical application

As a numerical example we discuss the prediction of games in the recent 2016 UEFA European Championship. Thus we follow Dyte and Clark (2000) and Suzuki et al. (2010) in considering applications to recent international tournaments. Historical records show that prior to the 2016 championship there had been an average of 2.46 goals per game in European Championship finals matches. We use data from FIFA's Coca Cola world rankings (data correct as of June 2nd 2016) as a proxy measure of team quality to estimate the value of p in the above (see Table 2). This follows a similar approach using FIFA team ratings in Dyte and Clarke (2000) and also shares some similarities with classical models of competitive balance (Andreff and Szymanski, 2006).

England's elimination from EURO 2016 at the hands of Iceland has been described as a "national sporting embarrassment" (McNulty, 2016). However, how does this supposedly shock defeat compare to the output of our model? From Table 2 we take $\lambda = 2.46$ and that the relative probability of an England goal is given by $p = 1069/(1069 + 751) = 0.587$. From equation (1) the probability of various scores are shown below in Table 3. Results suggest that defeat to Iceland may not quite be the national embarrassment many perceive. The most probable match score prior to kick-off is 1-1. According to our model the probability of a draw after 90 minutes is 0.265, the probability of an England win is 0.470, and the probability of an Iceland win is 0.265. Using equations (4-5) the probability of an overall England win (including extra time and penalties) is estimated to be 0.616 and the probability of an overall Iceland win is estimated to

Division	Goals per game (λ)	League title winners	Automatic promotion	Promotion playoffs	Avoid relegation
Champ'ship	$\lambda = 2.42$	$q = 0.971$ MV Target=87 MVK Target=91	$q = 22/23$ MV Target=85 MVK Target=87	$q = 18/23$ MV Target=73 MVK Target=73	$q = 3/23$ MV Target=49 MVK Target=49
League 1	$\lambda = 2.62$	$q = 0.971$ MV Target=87 MVK Target=91	$q = 22/23$ MV Target=85 MVK Target=88	$q = 18/23$ MV Target=73 MVK Target=73	$q = 3/23$ MV Target=51 MVK Target=49
League 2	$\lambda = 2.67$	$q = 0.971$ MV Target=88 MVK Target=91	$q = 21/23$ MV Target=81 MVK Target=82	$q = 17/23$ MV Target=72 MVK Target=71	$q = 2/23$ MV Target=46 MVK Target=45
National League	$\lambda = 2.64$	$q = 0.971$ MV Target=88 MVK Target=91	$q = 0.971$ MV Target=88 MVK Target=91	$q = 20/23$ MV Target=78 MVK Target=78	$q = 4/23$ MV Target=51 MVK Target=52
National League North ($n = 22$)	$\lambda = 2.79$	$q = 0.969$ MV Target=81 MVK Target=84	$q = 0.969$ MV Target=81 MVK Target=84	$q = 17/21$ MV Target=69 MVK Target=69	$q = 3/21$ MV Target=45 MVK Target=45
National League South ($n = 22$)	$\lambda = 2.92$	$q = 0.969$ MV Target=81 MVK Target=85	$q = 0.969$ MV Target=81 MVK Target=85	$q = 17/21$ MV Target=69 MVK Target=69	$q = 3/21$ MV Target=45 MVK Target=45
Isthmian Premier League	$\lambda = 3.07$	$q = 0.971$ MV Target=89 MVK Target=92	$q = 0.971$ MV Target=89 MVK Target=92	$q = 20/23$ MV Target=79 MVK Target=79	$q = 4/23$ MV Target=52 MVK Target=52
Northern Premier League	$\lambda = 3.02$	$q = 0.971$ MV Target=89 MVK Target=92	$q = 0.971$ MV Target=89 MVK Target=92	$q = 20/23$ MV Target=78 MVK Target=79	$q = 4/23$ MV Target=52 MVK Target=52
Southern Premier League	$\lambda = 2.87$	$q = 0.971$ MV Target= MVK Target=92	$q = 0.971$ MV Target=88 MVK Target=92	$q = 20/23$ MV Target=78 MVK Target=78	$q = 4/23$ MV Target=52 MVK Target=52

Table 1: English football league titles and suggested points targets for the mean-variance (MV) and mean-variance-kurtosis (MVK) approximations ($n = 24$ teams unless stated). Average number of goals, league structures and performance targets taken from the last fully completed league season (2015-2016).

be 0.384. Relatedly, results in Table 4 also suggest that Wales win over Belgium in the Quarter Finals of Euro 2016 was perhaps also more likely than bookmakers (and others) had anticipated.

In the sequel we compare the predictions of our model with bookmaker predictions for each of the EURO 2016 quarter final matches. First, we need to take into account the effects of home advantage. Academic work (Clarke and Norman, 1995) and commercial betting models suggest that in soccer home advantage equates to a net benefit of around 0.5 goals to the home team

Team	World Ranking	Points Total
Belgium	2	1384
Wales	26	846
Portugal	8	1181
Poland	27	842
Italy	12	982
Germany	4	1310
France	17	925
Iceland	34	751
England	11	1069

Table 2: FIFA World Rankings and points totals (proxy for team strengths) for the last nine teams to be eliminated from the 2016 UEFA European Championships.

Goals scored	Iceland 0	Iceland 1	Iceland 2	Iceland 3
England 0	0.085	0.087	0.044	0.015
England 1	0.123	0.125	0.064	0.022
England 2	0.089	0.091	0.046	0.016
England 3	0.043	0.044	0.022	0.007

Table 3: Probability of various scores for England v. Iceland in EURO 2016

compared to playing away from home. This suggests that playing at home gives a net advantage of around 0.25 goals to the home team compared to a match at a neutral venue. Thus, we assume that playing at home increases the expected number of goals scored by the home team by 0.125 and simultaneously reduces the number of goals scored by the away team by 0.125. In order to retain appropriately normalised probabilities this suggests the adjustments

$$\begin{aligned} \lambda p_{home} &= \lambda p + 0.125;\ p_{home} = p + \frac{0.125}{\lambda}; \\ \lambda p_{away} &= \lambda p - 0.125;\ p_{away} = p - \frac{0.125}{\lambda} \end{aligned} \tag{12}$$

Incorporating the correction for home advantage suggested by (12) gives that the size of the adjustment for the home nation France is $0.125/2.46 = 0.051$. Using equation (4-5) we compare the results of our model with implied market probabilities obtained from bookmakers' odds. The results are shown below in Table 4 and show reasonable agreement with market probabilities. However, market probabilities arguably over-estimate the probability that hot favourites such as Belgium and France will win their respective quarter-final matches. Note that this feature

runs counter to the so-called favourite-longshot bias (see e.g. Cain et al., 2000).

Match outcome	Poland win	Portugal win
Model Probability	0.389	0.611
Market Probability	0.406	0.594
Match outcome	Wales win	Belgium win
Model Probability	0.342	0.658
Market Probability	0.298	0.702
Match outcome	Germany win	Italy win
Model Probability	0.595	0.405
Market Probability	0.602	0.398
Match outcome	France win	Iceland win
Model Probability	0.636	0.364
Market Probability	0.792	0.208

Table 4: Model probabilities for EURO 2016 quarter finals compared to implied probabilities for bookmakers' odds.

Conclusions and further work

Though observed scores typically show small systematic deviations from a Poisson distribution (Boshnakov et al., 2015) Poisson modelling both offers a reasonable description of match outcomes and is well-established (Dyte and Clark, 2000; Suzuki et al., 2010). In particular, by placing constraints on the expected total number of goals in a match, which reflects known features of real soccer competitions, we can produce simple analytical formulae for match outcomes. In empirical applications further modifications are possible to variously correct for differing team strengths, the current score, home advantage and extra-time and penalties. Our model also provides a theoretical explanation of the empirically observed inverse strike-back effect. We thus provide a mathematical treatment of a serious practical problem. The elegance of our approach is underscored by the realistic performance targets that can be constructed under the simplifying assumption of perfect competition. In particular, our model reproduces versions of well-known rules of thumb that in order to avoid relegation from the English Premier League and the English Championship teams need to target around 41 points and 49 points respectively. Similar comments also apply to other leagues around the world such as Ligue 1 in France. A Mean-Variance-Kurtosis approximation is shown to offer marked improvements over a simple

Mean-Variance approximation when calculating performance targets.

In this paper we have explored the hidden implications of the classical Poisson model for soccer. Further work will extend will extend this model of perfect competition in Section 3 to other, more complicated, sports such as rugby and cricket. Amid much interest in sports modelling (see e.g. Wright, 2009) future work will develop analytical models with non-constant goal-scoring rates and adjusting for other important factors such as injuries, loss of form, psychology, fatigue etc. – see e.g. related applied work in Owen (2011) and Constantinou et al. (2012).

Appendix: The proofs

Proof of Proposition 3

Draws occur if both teams score n goals. The probability of a draw can thus be calculated as

$$\begin{aligned}\sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^{2n}}{2n!}.\frac{2n!p^n(1-p)^n}{n!n!} &= e^{-\lambda}\sum_{n=0}^{\infty}\frac{\left(\lambda\sqrt{p(1-p)}\right)^{2n}}{n!n!} \\ &= e^{-\lambda}I_0(2\lambda\sqrt{p(1-p)})\end{aligned}$$

The probability that team X wins by a margin of r goals can be calculated as

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^{2k+r}}{(2k+r)!}.\frac{(2k+r)!p^{k+r}(1-p)^k}{(k+r)!k!} = e^{-\lambda}\left(\frac{p}{1-p}\right)^{\frac{r}{2}} I_r\left(2\lambda\sqrt{p(1-p)}\right). \tag{13}$$

The probability that Team X wins can then be obtained by summing equation (13) over r to obtain

$$\begin{aligned}Pr(\text{Team } X \text{ wins}) &= e^{-\lambda}\sum_{r=1}^{\infty}\left(\frac{p}{1-p}\right)^{\frac{r}{2}} I_r(2\lambda\sqrt{p(1-p)}) \\ &= e^{-\lambda}e^{\lambda}Q_0(\sqrt{2\lambda p}, \sqrt{2\lambda(1-p)}),\end{aligned}$$

using

$$Q_0(\alpha,\beta) = e^{-\frac{\alpha^2+\beta^2}{2}}\sum_{k=1}^{\infty}\left(\frac{\alpha}{\beta}\right)^k I_k(\alpha\beta),$$

(see e.g. Proakis, 1983). □

Proof of Proposition 4

Parts (i-ii). From Proposition 3 the probability X wins outright in extra-time is given by

$$Pr(\text{X wins outright in extra-time}) = Q_0(\sqrt{2/3\lambda p}, \sqrt{2/3\lambda(1-p)}). \tag{14}$$

From equation (14) it follows that $Pr(X$ wins on penalties) is given by

$$\frac{1}{2}\left[1 - Q_0(\sqrt{2/3\lambda p}, \sqrt{2/3\lambda(1-p)}) - Q_0(\sqrt{2/3\lambda(1-p)}, \sqrt{2/3\lambda p})\right]. \tag{15}$$

Equation (2) follows by summing equations (14-15). Parts (iii-iv) follow since

$$Pr(\text{X wins}) = Pr(\text{X wins outright after 90 minutes}) + Pr(\text{X wins after extra-time})Pr(\text{Draw}).$$

□

Proof of Proposition 5

Suppose that a goal is scored at time M. The probability that the same team scores again next is given by

$$\left[p^2 + (1-p)^2\right]\left[1 - e^{-\frac{\lambda[90-M]}{90}}\right]. \tag{16}$$

The probability that the other team scores again next is given by

$$[2p(1-p)]\left[1 - e^{-\frac{\lambda[90-M]}{90}}\right]. \tag{17}$$

The probability in equation (16) is higher than in equation (17) since

$$4p - 4p^2 \leq 1;\ 4p - 2p^2 \leq 1 + 2p^2;\ 2p - 2p^2 \leq 1 + 2p^2 - 2p.$$

□

Proof of Proposition 6

Equation (10) is a mean-variance approximation. The mean follows from equation (7). The variance can be calculated using

$$\begin{aligned}\mathrm{Var}(X-Y) &= \sum_{i=1}^{2n-4}\mathrm{Var}(X_i-Y_i)+2\mathrm{Var}(Z_i) = (2n-4)2(5w-w^2)+2(18w)\\ &= [20n-4]w+w^2[8-4n].\end{aligned}$$

In the sequel we consider a higher-order Cornish-Fisher approximation (Fisher and Cornish, 1960). The form of equation (11) simplifies since from (8) the random variable $X-Y$ is symmetric and so has skewness zero. The excess kurtosis of the random variable $X-Y$ can be calculated as follows. Let $W_i = X_i - Y_i$ and let Z_i be defined as above with $\mathrm{Var}(W_i)=\sigma_1^2$, $\mathrm{Var}(Z_i)=\sigma_2^2$. Since $E[W_i]=E[Z_i]=0$ the Excess Kurtosis (Ex Kurt) of $X-Y$ can be calculated as

$$\begin{aligned}\text{Ex Kurt}(X-Y) &= \frac{E[(W_1+\ldots+W_{2n-4}+Z_{2n-3}+Z_{2n-2})^4]}{\mathrm{Var}(W_1+\ldots+W_{2n-4}+Z_{2n-3}+Z_{2n-2})}-3\\ &= \frac{(2n-4)E[W_i^4]+2E[Z_i^4]+6\left(\frac{(2n-2)(2n-3)}{2}\sigma_1^4+2(2n-4)\sigma_1^2\sigma_2^2+\sigma_2^4\right)}{((2n-4)\sigma_1^2+2\sigma_2^2)^2}-3\\ &= \frac{(2n-4)E[W_i^4]+2E[Z_i^4]+\sigma_1^4[18n-30]-6\sigma_2^4}{((2n-4)\sigma_1^2+2\sigma_2^2)^2}. \qquad (18)\end{aligned}$$

We have that $\sigma_1^2=10w(1+w)$, $\sigma_2^2=18w$, $E[Z_i^4]=162w$. Further, $E[W_i^4]=E[(X_i-Y_i)^4]=94w^2+34w$. Inputting these values into (18) the stated result follows. □

References

[1] Abramowitz M and Stegun I (eds) (1968). Handbook of mathematical functions. Dover: New York.

[2] Andreff W and Szymanski S (eds) (2006) Handbook on the exonomics of sport. Edward Elgar: Cheltenham Northampton, Massachussets.

[3] Anderson C and Sally D (2013). The numbers game: Why everything you know about football is wrong. Penguin: London.

[4] Annamalai, A. and Tellambura, C. (2008) A simple exponential integral representation of the Generalized Marcum Q-Function $Q_M(a, b)$ for real-order M with applications. In Milcom 2008 - 2008 IEEE Military Communications Conference pp. 1-7.

[5] Baker R D and McHale I G (2015). Time varying ratings in association football: the all time greatest team is ... Journal of the Royal Statistical Society A178: 481-492.

[6] Boshnakov G, Kharrat T and McHale I G (2015). Are goals Poisson distributed? STN Journal of Sports Modelling and Trading (forthcoming)

[7] Cain L P and Haddock D D (2006). Measuring parity: Tying into the idealized standard deviation. Journal of Sports Economics 7: 330-338.

[8] Cain M, Law D and Peel D (2000). The favourite-longshot bias and market efficiency in UK football betting. Scottish Journal of Political Economy 47: 25-36.

[9] Campbell J Y Lo A and MacKinlay A C (1997) The econometrics of financial markets, Princeton University Press: Princeton.

[10] Clarke S T and Norman J M (1995)Home ground advantage of individual clubs in English soccer. Journal of the Royal Statistical Society D 44: 509-526.

[11] Constantinou A C and Fenton N E (2017) Towards smart-data: Improving predictive accuracy in long-term football team performance. Knowledge-Based Systems 124: 93-104.

[12] Constantinou, A C., Fenton N E and Neil, M. (2012) pi-football. A Bayesian network model for forecasting Association football match outcomes. Knowledge-Based Systems 36: 322-329.

[13] Dixon, M J and Robinson M E (1998). A birth process model for association football matches. Journal of the Royal Statistical society D 47: 523-538.

[14] Dyte D and Clarke S R (2000). A ratings based Poisson model for World Cup soccer simulation. Journal of the Operational Research Society 51: 993-998.

[15] Espitia-Escuer M and García-Cebrián L I (2010). Measurement of the efficiency of football teams in the Champions League. Managerial and Decision Economics 31: 373-386.

[16] Fisher R A and Cornish E A (1960). The percentile points of distrbutions having known cumulants. Technometrics 2: 209-225.

[17] Fitt A D, Howls C J and Kabelka M (2006). Valuation of soocer spread bets. Journal of the Operational Research Society 57: 975-985.

[18] Gandy R (2016). Second season syndrome. Significance 13: 26-29.

[19] Hyndman R J and Fan Y (1996) Sample quantiles in statistical packages. The American Statistician 50: 361-365.

[20] Jessop A (2006). A measure of compeitiveness in leagues: a network approach. Journal of the Operational Research Society 57: 1425-1434.

[21] Keller J B (1994). A characterisation of the Poisson distribution and the probability of winning a game. The American Statistician 48: 294-298

[22] Kuper S and Szymanski S (2014). Soccernomics. HarperSport: London.

[23] Lee R and Fort Y H (2012). Competitive balance: Time series lessons from the English Premier League. Scottish Journal of Political Economy: 59 266-282

[24] Maher M J (1982). Modelling association football scores. Statistica Neerlandica 36: 109-118.

[25] McNulty P (2016). Euro 2106: FA and England will sift through wreckage of embarrassment. BBC website. `http://www.bbc.co.uk/sport/football/36647964`

[26] Nuttall A H (1975). Some integrals involving the Q_M function. IEEE Transactions on Information Theory 21: 95-96.

[27] Owen A (2011). Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution parameter. IMA Journal of Management Mathematics 22: 99-113.

[28] Proakis J (1983). Digital Communications. McGraw-Hill: New York.

[29] Percy, D F (2015). Strategy selection and outcome prediction in sport using dynamic learning for stochastic processes. Journal of the Operational Research Society 66, 1840-1849.

[30] Rottenberg S (1956). The baseball players' labour market. Journal of Political Economy 64: 242-258.

[31] Suzuki A K, Salasat, L E B, Leite J G and Louzado-Neto, F (2010). A Bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup. Journal of the Operational Research Society 61: 1530-1539.

[32] Weingberg G V (2006). Poisson representation and Monte Carlo estimatin of Generalized Marcum Q-Function. IEEE Transactions on Aerospace and Electronic Systems 42: 1520-1531.

[33] Wright M B (2009). 50 years of OR in sport. Journal of the Operational Research Society 60 5161-5168.